Policy Research Working Paper 5628

Measuring True Sales and Underreporting with Matched Firm-Level Survey and Tax-Office Data

Fujin Zhou
Remco Oostendorp

The World Bank
East Asia and Pacific Region
Poverty Reduction and Economic Management Unit
April 2011

Policy Research Working Paper 5628

Abstract

This paper uses firm-level survey data matched with official tax records to estimate the unobserved true sales of formal firms in Mongolia. Taking into account firm-level incentives to comply with taxes and a production function technology linking unobserved true sales with observable firm-level production characteristics, the authors derive a multiple-indicators, multiple-causes model predicting unobserved true sales. Comparing predicted true sales with sales reported to the tax office, the analysis finds that 38.6 percent of firm-level sales are underreported. It also finds evidence that firm-level survey data suffer from significant underreporting. Finally, the paper compares this approach with two alternative approaches to measuring underreporting by firms.

This paper is a product of the Poverty Reduction and Economic Management Unit, East Asia and Pacific Region. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org.
The authors may be contacted at f.zhou@vu.nl and r.oostendorp@vu.nl.

Measuring True Sales and Underreporting with Matched Firm-Level Survey and Tax-Office Data[1]

Fujin Zhou[2]

Remco Oostendorp[3]

Key words: MIMIC, underreporting, investment climate, tax compliance model

JEL: C39, D24, E26, H26

[1] The authors gratefully acknowledge the valuable comments from Eric Bartelsman, Chris Elbers, and Erik Plug.
[2] Tinbergen Institute Amsterdam, and VU University Amsterdam. Email: *f.zhou@ vu.nl*.
[3] VU University Amsterdam , Tinbergen Institute Amsterdam, and Amsterdam Institute for International Development. Email: *r.oostendorp@ vu.nl*.

1 Introduction

"All Cretans are liars." (*Epimenides, Philosopher from Knossos, Crete, circa 600 BC)*

Over the past decades, firm-level data have become widely available for economic research and are used in manifold and vibrant lines of research. This is especially true for research on firm behavior in developing countries. The World Bank 'Regional Program on Enterprise Development' (RPED) was the first large-scale effort to gather firm-level survey data in developing countries using a multipurpose survey instrument. The survey was executed in eight Sub-Saharan countries in the early 1990s and the resulting survey data have been extensively used in research on firm investment, export behavior, rent- and risk-sharing (Teal 1996, Bigsten, et al. 2003). With the success of the RPED surveys, more large-scale firm-level surveys in developing countries followed, such as the 'World Business Environment Surveys' (WBES), '(Productivity and) Investment Climate Surveys' ((P)ICS), and the 'Business Environment and Enterprise Performance Surveys' (BEEPS), organized by the World Bank, European Bank for Reconstruction and Development (EBRD) and/or other (multilateral) institutions. Also these recent firm-level surveys have attracted the attention of economic researchers and have been analyzed intensively (Dollar, Hallward-Driemeier and Mengistae 2005, Cull and Xu 2005). [4]

Of course, researchers have been well aware that firms may have an incentive to misreport their activities, for reasons such as high marginal tax rates, corruption, and crime (Johnson, et al. 2000, Dabla-Norris and Koeda 2008, Gatti and Honorati 2008). Firm-level studies on underreporting of firm sales and/or output often rely on self-reported measures of

[2] A comprehensive review of the literature using firm-level data from developing countries is beyond the scope of this paper.

underreporting by firms, formulated in terms of the typical behavior of a firm in the same area of activity. The fact that these studies often report a sizeable degree of underreporting together with intuitive correlations with observable firm and investment climate characteristics suggests that these self-reported measures capture underreporting to some degree. However, it remains unclear how reliable these measures are without further probing the underlying assumption that firms report truthfully about untruthful reporting (sic).[5]

Firm-level studies which do not specifically focus on informality and/or misreporting almost always assume that firms report truthfully or that firm-level measures suffer from classical measurement error only. However, to the extent that misreporting behavior is systematically related with (observable and/or unobservable) firm-level and investment climate characteristics for which the analysis does not control adequately, the reported results will suffer from systematic (and unknown) measurement error bias. Also if one relies on survey rather than tax office data in the analysis, it is not clear to which extent the survey data suffer less from misreporting than tax office data.

Using unique firm-level survey data matched with official tax data, we attempt to estimate the unobserved true sales and the underreporting in sales to the tax office of formal sector firms in Mongolia. Based on the existing approaches used in the shadow economy literature, we can distinguish among three possible ways of estimating this underreporting.[6]

[5] In this respect it is interesting to note that at the time the initial RPED surveys were planned serious doubts were raised whether reliable data could be generated at all with structured questionnaires in a large-scale survey, especially in developing countries.
[6] Schneider and Enste (2000) provide a comprehensive review of the three approaches.

The direct approach is a micro approach that uses surveys to reveal the extent of underreporting directly. Respondents are either randomly sampled or selected as part of tax auditing or other compliance methods. For instance, the (P)ICS and BEEPS surveys include a question which is typically formulated along the following lines: "Recognizing the difficulties many firms face in fully complying with taxes and regulations, what percentage of total annual sales would you estimate the typical firm in your area of business reports for tax purposes?" Together with other information collected on respondents' behavior and environment, these sample surveys provide rich information about underreporting and its correlates but are sensitive to the formulation of questionnaires and largely depend on the respondents' willingness to cooperate. Tax auditing methods may be better able to extract truthful information from the auditees, but tax authorities may not be able to fully discover the true incomes of the audited group. Moreover, the audited group is typically a biased sample of the population (Schneider and Enste 2000).

The second type of approach is the indirect approach or indicator approach and has been primarily used in macroeconomic settings. The approach consists of constructing indicators that reflect the development of a shadow economy over time, such as the discrepancy between national expenditure and income statistics and/or between the official and actual labor force. The quality of the approach depends therefore on the accuracy of these indicators and may be seriously affected by measurement errors and systematic underreporting in the indicators (Giles 1999, Schneider and Enste 2000).

The third approach to estimating the extent of underreporting is the model approach. This approach was introduced into economics by Frey and Weck-Hannemann (1984) in their

study of the size of the hidden economy of a cross-section of 17 OECD countries for the period 1960-1978 and has been used in several studies thereafter (Loayza 1996, Chaudhuri, Schneider and Chattopadhyay 2006, Dell'Anno, Gomez-Antonio and Pardo 2007). The approach is based on structural equation modeling with latent variable(s), for which multiple causes and multiple indicators exist (MIMIC model). The empirical MIMIC literature is primarily macroeconomic in nature and typically the specification of the applied MIMIC model is not derived from formal economic theory. One notable exception is a paper by Siegel (1997), which uses a MIMIC model derived from formal economic theory to estimate the contribution of computer usage in productivity growth at the industry level.

This paper makes a number of contributions. First, because we have unique firm-level survey data matched with official tax data, we can apply and compare the above three approaches simultaneously for the same sample of firms. Second, unlike most of the MIMIC literature, we use formal economic theory to derive a MIMIC model to estimate the size of hidden outputs using microeconomic data. In particular, we model true sales by taking into account firm-level incentives to comply with taxes and a production function technology linking true sales with observable firm-level production characteristics. Third, we not only allow for underreporting in the official tax office data but also for underreporting in the survey data. Fourth, we estimate the extent of underreporting for a sample of formal sector firms in the transition economy of Mongolia where the extent of underreporting is expected to be prevalent but largely unknown.[7] The finding for Mongolia should also be

[7] While not studying underreporting by formal sector firms specifically, a number of studies have looked at the broader issue of shadow economy in Mongolia before (Anderson 1998, Bikales, Khurelbaatar and Schelzig 2000).

relevant for many other developing (transition and non-transition) countries where we may expect serious underreporting.

We will argue that the MIMIC approach provides the more accurate estimate of the extent of underreporting because it incorporates firm-level incentives to comply with taxes and a production function technology linking true sales with observable firm-level production characteristics, it controls for measurement errors, and it allows for underreporting in both official tax and survey data. According to the MIMIC approach, the average percentage of underreporting to the tax office is 38.6% at the firm-level and 37.5% at the aggregate for the population of firms from which the sample has been drawn. The indirect approach performs poorly and underestimates underreporting because it is sensitive to measurement errors and underreporting in the survey data. The direct approach gives an estimate of the firm-level average percentage of underreporting which is somewhat lower than the MIMIC approach, confirming the conjecture of Schneider and Enste (2000, p.92) that the direct approach provides a lower bound for true underreporting. However, the direct approach gives a too high estimate for aggregate underreporting because of measurement error and appears less useful as an indicator of underreporting by individual firms.

The remainder of the paper is organized as follows. In section 2 we discuss the extent of underreporting by firms around the world based on the commonly used self-reported measures of underreporting (direct approach). In section 3 we provide a formal derivation of a MIMIC model to estimate the true sales of a firm. In section 4 we discuss the empirical

specification of the model and present the empirical results. Section 5 concludes with a comparison of the MIMIC results with those from direct and indirect approaches.

2 Underreporting by firms around the world – the direct approach

While most of the existing empirical research on the unofficial economy uses macro data, a number of recent papers have used firm-level survey data to analyze the determinants and consequences of underreporting of sales and/or output by firms (Johnson, et al. 2000, Gatti and Honorati 2008). These papers take advantage of the increasing availability of large-scale and comparable firm-level surveys in which firms have been asked about their tax reporting behavior, such as in the BEEPS, (P)ICS and WBES surveys. Because the firms in these surveys are typically registered firms, the papers analyze informal activity by otherwise formal (registered) firms using the so-called direct approach.

Because of the sensitive nature of the subject, the question about reporting behavior is phrased in terms of actions of 'typical firms in your area of activity': *Recognizing the difficulties many enterprises face in fully complying with taxes and regulations, what percentage of total sales would you estimate the typical establishment in your area of activity reports for tax purposes?* Researchers have explicitly or implicitly interpreted the firms' responses to this question as indicators of underreporting at the firm-level.

These studies often report a sizeable degree of underreporting together with intuitive correlations with observable firm and investment climate characteristics. This suggests that these self-reported measures do capture underreporting, at least to some degree. Moreover, Johnson *et al*. (2000) also note that the estimates of hidden activity from the

surveys are quite similar to those available from two independent 'macro' methodologies.[8] This suggests that the direct approach based on surveys to measuring unofficial activity provides useful information about actual underreporting behavior by firms.

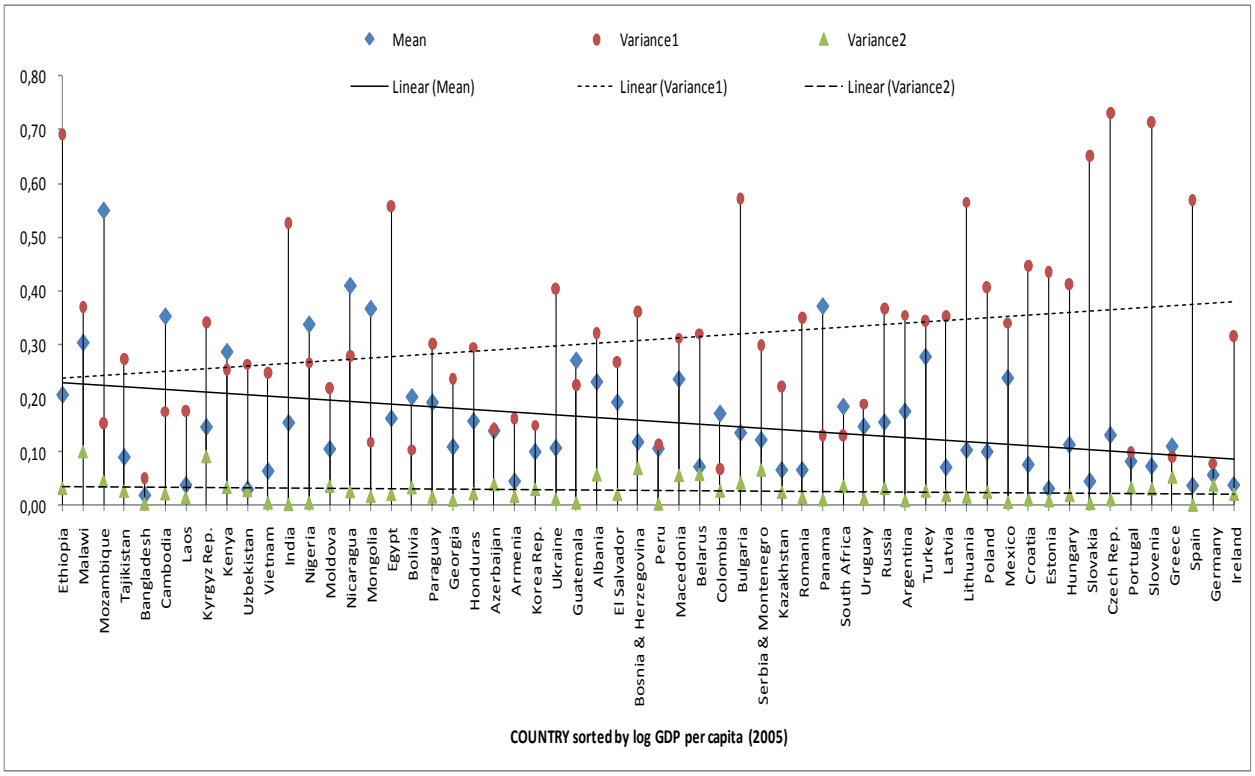
In figure 1 we plot the mean of underreporting by firms based on the direct approach across countries sorted by (log) GDP per capita in 2005. We used all BEEPS, (P)ICS, WBES and Firm Analysis and Competitiveness Surveys (FACS) for 2005 or later that could be freely accessed through the World Bank website[9] and which included information on self-reported underreporting in sales. In case multiple surveys were available for a given country, we selected the most recent one. Figure 1 shows a clear pattern that the mean level of underreporting is decreasing in log of GDP per capita. Underreporting is the highest across the low income countries and the lowest across the high income and OECD countries[10]. However, even at comparable levels of income, there is a large variation in the extent of underreporting by firms across countries, which may reflect genuine country-differences in underreporting behavior as well as measurement errors.

Because the survey asks about the underreporting by typical firms in the same area of activity, there may be a "bias towards the average behavior of other firms in that environment" (Dabla-Norris and Koeda 2008, p.10) as firms may report the average amount of underreporting for firms in the same industry. Also they may report the average amount of underreporting as perceived by them in the same locality. We therefore

[8] Although one should note that the macro estimates are for the percent of GDP that is unregistered, rather than the percent sales not reported by formal sector firms to the tax authorities.
[9] www.enterprisesurveys.org
[10] Based on the World Bank classification of income groups.

calculated the percent of the variance in reported underreporting that can be explained by location and industry effects ('Variance 1' in Figure 1). If firms report the average amount of underreporting for firms in the same industry and/or locality, then a major part of the

Figure 1: Underreporting by Firms around the World Measured by Direct Approach



observed variance in reported underreporting would be captured by location, industry and interaction effects. For the low income countries the percentage explained is 26.1% on average, for the lower and upper middle income countries this is 27.6% and 28.7% respectively, and for the high income and OECD countries they are 56.5% and 23% respectively. Hence, it appears that a major part of the variance in reported underreporting can be explained by differences in mean reporting of firms active in different industries

and/or localities.[11] On the other hand, only a small part of the variance appears to be correlated with firm-level characteristics once we control for industry and location-specific effects. 'Variance2' in figure 1 is the additional variance that can be explained by firm size and ownership dummies. Ownership is captured by 5 ownership categories and firm size by 3 size categories in all countries. Using more size categories or a polynomial function of a continuous size (i.e. employment) variable does not increase the variance explained much. The figure shows that the firm-level characteristics firm size and ownership explains a small percent of the observed variance once we control for the variation across industries and locations. 'Variance 1' is uniformly much larger than 'Variance 2'. Although one might also include other firm-level characteristics besides firm size and ownership, there are no obvious other firm-level variables that will be able to explain much more of the total variance.

Therefore the largest part of the explained variance in reported underreporting is indeed across industries and localities, rather than across firms within industries and localities. This suggests that the responses are indeed biased towards the average behavior of other firms in the environment making them less indicative of underreporting by individual firms.[12]

[11] The relatively low percentages explained for Colombia, Germany, Greece, and Portugal are probably due to the fact that in these surveys individual cities are aggregated into size classes.

[12] Unless one assumes that underreporting behavior is mostly random and uncorrelated with observable firm characteristics.

There remains, however, the important issue of whether the direct approach provides reliable information, even if only about average underreporting behavior. It has been noted that the direct approach may provide a lower-bound estimate of actual underreporting (Schneider and Enste 2000, 92). The direct approach relies on the untested assumption that firms report truthfully about untruthful reporting, but this may not hold in practice. Therefore in the next section we develop an alternative approach which allows for (systematic and nonsystematic) misreporting in the survey and which estimates underreporting at the firm-level.

3 A MIMIC model

Our approach is to derive a MIMIC model for the estimation of the true sales of a firm taking into account firm-level incentives to comply with taxes and a production function technology linking true sales with observable firm-level production characteristics. Generally, a MIMIC model is a structural equation model with latent variable(s) (y^*) for which multiple indicators (Y) and multiple causes exist (X). Apart from the shadow economy literature mentioned in the introduction, the MIMIC model has also been applied to estimate the demand for health care (Van de Ven and Van der Graag 1982) and manufacturing productivity growth (Siegel 1997). The standard MIMIC model consists of both structural equation(s) (eq. 1) and measurement equation(s) (eq. 2):

$$y^* = \alpha' X + \varepsilon \tag{1}$$

$$Y = \beta y^* + u \tag{2}$$

with vectors $X = (x_1 \dots x_k)'$, $\alpha = (\alpha_1 \dots \alpha_k)'$, $Y = (y_1 \dots y_m)'$, $\beta = (\beta_1 \dots \beta_m)'$, $u = (u_1 \dots u_k)'$. In our case, the latent variable is (unobserved) true sales, which depends on a number of causes or factors X (the structural equation 1), and which is measured with error through indicators Y (the measurement equation 2). In the standard MIMIC model, it is also assumed that $E(\varepsilon\varepsilon') = \sigma^2, E(uu') = \Theta^2, E(\varepsilon u') = 0$, with Θ^2 a diagonal covariance matrix. If ε and u are jointly normally distributed, maximum likelihood estimation can be applied to the reduced form of the MIMIC model to obtain parameter estimates. [13]

Our MIMIC model will be a more general version of the standard MIMIC model. First, we will allow for correlation between the error terms ε and u ($E(\varepsilon u') \neq 0$), and between $u_1 \dots u_k$ (Θ^2 non-diagonal). Second, we allow indicators not only to depend on the latent variable but also on other factors Z as well ($Y = \beta y^* + \gamma'' Z + u$). And third, we will allow for truncation given that reported sales to the tax office (an indicator in our model) should not exceed the true sales of the firm (the latent variable). We first discuss the derivation of our structural equation and next continue with a discussion of our derivation of two measurement equations for the MIMIC model.

A. Structural Equation: Production Technology

Assume that the production technology of a firm can be represented by a Cobb-Douglas production function:

$$Y = AM^{\alpha}L^{\beta}K^{\gamma} \tag{3}$$

[13] For more details about the identification and estimation of the MIMIC model, see Joreskog and Goldberger (1975).

where Y is the quantity of output, M, L and K are raw material, labor and capital inputs respectively, and A is total factor productivity; α, β and γ is the elasticity of production with respect to M, L and K in the output function. Let p, p_M, w and r denote the price of output, the price of raw material inputs, the wage rate and the user cost of capital respectively. We rewrite the production function in terms of values since in our empirical analysis only data on the value of sales, raw materials, and capital is available. Further taking the natural logarithmic transformation, we obtain:

$$\ln S = \ln \mu + \alpha \ln M_v + \beta \ln L_v + \gamma \ln K_v \quad (4)$$

where $\mu = \frac{pA}{p_M^{\alpha} w^{\beta} r^{\gamma}}$, $S = pY$ is the value of sales[14] and $M_v = p_M M$, $L_v = wL$, $K_v = rK$ denote the values of raw materials, labor costs, and capital respectively. Hence, the intercept $\ln \mu$ reflects the impact of prices and productivity on the value of output.

In the short run, capital is fixed and firms maximize profits over the choices of raw material and labor inputs, taking input prices (which could be firm-specific) as given. If we substitute the profit-maximizing levels of M and L into (4), we obtain an expression for the optimal (log) sales (S^*):

$$\ln(S^*) = \frac{\alpha \ln \alpha + \beta \ln \beta}{1-\alpha-\beta} + \frac{\ln \mu}{1-\alpha-\beta} + \frac{\gamma \ln K_v}{1-\alpha-\beta} \quad (5)$$

The capital service flow variable K_V is taken as a weighted average of the values of capital stocks such as machinery and equipment, vehicles, and buildings (Christensen and Jorgensen 1969):

[14] We assume firms do not have inventory so that output values are equal to sales values.

$$K_v = (r + \delta_B)K_B + (r + \delta_{ME})K_{ME} \quad (6)$$

where K_B and δ_B are the stock values of and the depreciation rate for buildings; K_{ME} and δ_{ME} are the stock values of and the depreciation rate for machinery, equipment and vehicles; r is the return rate (or rental price) of capital.[15]

The term $\ln\mu$ in equation (5) is not observed without firm-level information on total labor productivity and prices. Price and productivity dispersion have been well documented in the literature and therefore $\ln\mu$ will vary across firms (Bartelsman and Doms 2000). Without direct measures of total factor productivity and prices, firm characteristics such as the location and industry of firms, firm size, ownership structure, and the investment climate faced by firms (e.g. corruption and regulation burden) are used to approximate the ratio $\ln\mu$. If we denote these firm-level observable proxy variables by the vector X_1, equation (5) can be written as:

$$\ln(S^*) = a + X_1 b + c \ln K_v + u \quad (7)$$

where a and b are parameters and u is the disturbance capturing either measurement errors or productivity shocks or other unobserved effects that are not captured by X_1. Equation (7) is a structural equation for unobserved ('true') sales (latent variable) and can be viewed as a short-run supply function.

[15] Hours worked per capital are a good proxy for capital service flow, but we do not have such data.

B. First Measurement Equation for Tax Office Data : Tax Compliance Model

There are two available types of indicators for the latent variable true sales, one from the official tax data and one from the firm-level survey data. Both indicators are required for identification purpose (see below). Reported sales to the tax office is a potential valuable indicator for unobserved true sales. However, it does not form an unbiased estimator of true sales because there are strong incentives for firms to reduce the tax burden by underreporting.

Allingham and Sandmo (1972) is one of the earliest and best known models of tax evasion, in which the individual taxpayer's decision on the level of tax compliance is subject to an exogenous and positive audit probability and a penalty when evasion is detected. Later extensions of this tax evasion model include endogenous labor supply decisions and audit probabilities, allowances for repeated interactions between tax payer and tax office, and taxpayer attitude and social dynamics (Reinganum and Wilde 1985, Graetz and Reinganum 1986, Grasmick and Bursick 1990, Erard and Feinstein 1994a). Also there is a small literature on corporate tax evasion (Crocker and Slemrod 2005, Chen and Chu 2005).

We formulate a simple tax compliance model to derive a relationship between the sales reported to the tax office and the true sales.[16] Firms have incentives to underreport taxable incomes through underreporting in sales and/or over-reporting in raw material costs. As for labor costs, over-reporting lowers taxable incomes but increases payroll taxes. So

[16] The model is constructed purely from a firm's perspective without modeling the strategic interaction between a firm and the tax authority because information on the tax authority is lacking.

whether to underreport or over-report labor costs depends on the relative cost of the income and payroll tax.

Tax evasion, however, is associated with uncertainty and incurs extra costs. For example, firms have to invest extra resources to make "double accounts"; under-the-counter transactions might hinder firms from fully utilizing public services such as legal and judicial systems, access to formal finance, et cetera. We assume the extra cost associated with tax evasion to be linearly dependent on the amount of profits underreported $(S^* - M_v^* - L_v^*)$ - $(S^t - M^t - L^t)$ with a coefficient δ, where S^*, M_v^*, L_v^* denote the true sales/raw materials/labor costs while S^t, M^t, L^t denote the reported sales/raw materials/labor costs to the tax office respectively. Therefore we assume that firms choose an optimal combination of sales, raw materials and labor costs reported to the tax office for retained profit maximization, taking into account the costs and benefits associated with tax evasion.

Since the tax authority conducts tax auditing with a budget constraint, only a small portion of the firms will be selected for tax audits. Firms know that the tax office faces a budget constraint and form a subjective (possibly firm-specific) perception of the efficiency of detecting underreporting by the tax office. Also they believe that the probability of detection increases with the extent of misreporting. More specifically, we assume that the subjective probabilities that the tax office will detect underreporting (or over-reporting) in sales, raw materials, and labor costs are respectively[17]:

[17] The three probabilities are mutually uncorrelated but in reality we expect them to be correlated. Our specification also restricts raw materials to be over-reported. However, we've verified that it is possible to relax both of the restrictions and the same implied relationship between true and reported values still holds under suitable assumptions.

$$\theta_S = \bar{\theta}\left|\frac{S^* - S^t}{S^*}\right|, \theta_M = \bar{\theta}\left|\frac{M_v^* - M^t}{M_v^*}\right|, \theta_L = \bar{\theta}\left|\frac{L_v^* - L^t}{L_v^*}\right|$$

where $\theta_i \in [0,1], i = S, M, L$ denote the detection probabilities for sales, raw materials, and labor costs respectively and $\bar{\theta}$ is the subjective expectation of tax office efficiency. For given tax office efficiency ($\bar{\theta}$), the detection probabilities increase with the relative gap between the true values and the reported values to tax office. We further assume that once a firm is chosen for tax auditing, any tax evasion will be detected and the true sales/raw materials/labor costs of the firm will be fully revealed.[18] Therefore, a firm maximizes the expected self-retained profit over the choices of reported sales, raw materials and labor costs, taking true sales/raw materials/labor costs as given. A firm's expected retained profit can be written as:

$$\max_{S^t, M^t, L^t} E(\pi) = [(S^* - M_v^* - L_v^*) - \tau\left(S^t - M^r - (1+\tau_L)L^t\right) - \tau_L L^t - (P\left(\theta_S|S^* - S^t| + \theta_M|M_v^* - M^t| + \theta_L|L_v^* - L^t|\right))] \\ -\delta((S^* - M_v^* - L_v^*)) - (S^t - M^t - L^t)) \quad (8)$$

where π denotes firm's retained profit, τ and τ_L the corporate profit tax rate and payroll tax rate respectively[19]; and P is the penalty multiplier. Plugging in the three audit probabilities into equation (8), we obtain three first-order conditions with respect to $\left(S^t, M^t, L^t\right)$:

$$(\mathbf{a})\ \frac{S^t}{S^*} = 1 - \frac{\tau - \delta}{2P\bar{\theta}}, (\mathbf{b})\ \frac{M^t}{M_v^*} = 1 + \frac{\tau - \delta}{2P\bar{\theta}}, (\mathbf{c})\ \frac{L^t}{L_v^*} = 1 + \frac{\tau - \delta - \tau_L(1-\tau)}{2P\bar{\theta}}$$

[18] We can also assume that only part of the underreporting will be detected but this does not change the main implications of the model.

[19] In 2003, Mongolia there were two profit tax rate levels for registered firms: 15% if the taxable income is below 100 million MNT and 40% above. We calculated the gross taxable incomes using tax office data and no firms have taxable incomes above the threshold of 100 million MNT. The calculated taxable incomes could be even lower if there are other tax deductible costs not included in the calculation. Hence assuming a single profit tax rate is reasonable. Also the payroll tax rate is flat –formal firms in Mongolia are obliged to pay 19% social security tax on wages or a slightly higher rate (or 20 and 21% in some industries with high injury risks).

Equations (a) to (c) show the relationship between the tax office sales/raw materials/labor costs and the true sales/raw materials/labor costs with equation (a) forming the analytical underpinning for a measurement equation for S^*. It's reasonable to expect that firms never report more than what they sell ($1 \geq \frac{S^t}{S^*} \geq 0$).[20] Therefore we can transform the first-order condition for sales into log linear form:

$$\ln S^t = \ln(S^*) + \ln\left(1 - \frac{\tau - \delta}{2P\bar{\theta}}\right) \tag{9}$$

The marginal effects of τ, P, δ and $\bar{\theta}$ on sales are:

$$\frac{\partial \frac{S^t}{S^*}}{\partial \tau} = \frac{-1}{(2P\bar{\theta})} < 0, \frac{\partial \frac{S^t}{S^*}}{\partial P} = \frac{\tau - \delta}{(2P^2\bar{\theta})} > 0, \frac{\partial \frac{S^t}{S^*}}{\partial \delta} = \frac{1}{(2P\bar{\theta})} > 0, \frac{\partial \frac{S^t}{S^*}}{\partial \bar{\theta}} = \frac{\tau - \delta}{(2P\bar{\theta}^2)} > 0$$

Reported sales increase with the penalty multiplier, probability of audit, detection efficiency, and the costs associated with misreporting, but decreases with corporate income tax rate. The patterns conform to our expectation and are in line with the predictions derived from other tax compliance models.

The values of the parameters P, δ and $\bar{\theta}$ are unknown, but can be approximated by a vector of observable proxy variables X_2 (see section 4). Observable proxy variables may also be used for the tax rate τ, if this is interpreted as an unobserved and firm-specific effective tax

[20] $\frac{S^t}{S^*} \leq 1$ holds if $\tau > \delta$, i.e. if the tax rate is higher than the marginal cost of tax evasion; and $\frac{S^t}{S^*} > 0$ holds as long as the firm perceived tax office efficiency $\bar{\theta}$ is not too small. $\frac{L^t}{L_v^*}$ can be larger than one if $\tau - \delta > \tau_L(1 - \tau)$ or smaller than one if $\tau - \delta < \tau_L(1 - \tau)$. Accordingly, firms may under-/over- report labor costs.

rate, rather than a nominal rate. Assuming a linear approximation for the unknown function $\ln\left(1-\frac{\tau-\delta}{2P\bar{\theta}}\right)$, we have:

$$\ln S^t = \ln(S^*) + d_1 + X_2\delta_2 + \varepsilon_t \quad (10)$$

where ε_t denotes all unobserved effects that are not captured by X_2 but which cause S^t to deviate from true sales. Equations (7) and (10) form a MIMIC model with one indicator and one latent variable. Plugging equation (7) into equation (10) we obtain a reduced form which can be estimated. However, the constants d_1 and a cannot be identified separately and parameter values in δ_2 and b for variables which occur in both X_1 and X_2 are also not identified. We need at least one more indicator and measurement equation in order to identify the key coefficients for predicting the latent variable true sales.

C. Second Measurement Equation for Survey Data: Measurement Error Model

The second indicator for the latent variable of true sales is from the 2003 Mongolian Productivity and Investment Climate Survey data organized by the World Bank.[21] Without an obvious economic theory to explain (mis)reporting behavior in the PICS survey, we assume a standard log linear measurement error model linking reported survey and true sales:

$$\ln S^s = d_2 + \beta_2 \ln(S^*) + \varepsilon_s \quad (11)$$

[21] See section 4 and Appendix A for more information on the survey.

where S^s is the sales reported in the survey, d_2 is a constant, and ε_s is the corresponding measurement error. The reported sales are expected to be positively correlated with true sales ($\beta_2 > 0$). We specify the survey sales to be a function of the true sales because at the time of the survey true sales are known to the firm. The log linear specification is flexible as it allows the survey sales to be either above or below the true sales.

Summarizing, equations (7), (10) and (11) form a MIMIC model, where a firm's true sales (S^*) is the latent variable and reported sales to the tax office (S^t) and in the survey (S^s) are the indicators (measures). We also include the restriction that reported sales to the tax office do not exceed the true sales (equation 12):

$$\ln(S^*) = a + X_1 b + c \ln K_v^* + u \quad (7)$$

$$\ln S^t = \ln(S^*) + d_1 + X_2 \delta_2 + \varepsilon_t \quad (10)$$

$$\ln S^s = d_2 + \beta_2 \ln(S^*) + \varepsilon_s \quad (11)$$

$$S^t \leq S^* \leftrightarrow \ln(S^t) \leq \ln(S^*) \quad (12)$$

In line with the MIMIC literature, we assume $u, \varepsilon_t, \varepsilon_s$ to be multivariate normally distributed with mean zero but we allow for any possible correlation among these error terms. In the Appendix B we discuss the identification of the parameters of the model and derive the likelihood function for estimation in the next section.

4 Estimated true sales versus reported sales

The data for this study are from Mongolia - a land-locked country in East and Central Asia that has gone through radical changes from central planning towards market economy in 1990s. We use two data sources. The first source of data is the World Bank Productivity

and Investment Climate Survey for Mongolia (PICS) from 2004. Data for 2002 and 2003 was collected but we focus on the 2003 data because data for several variables of interest are not available for 2002. The survey covers Mongolian registered firms with at least 3 employees from manufacturing, construction, service, and tourism sectors.[22] The coverage rates of the number of firms in the four sectors are 81%, 70%, 56% and 53% respectively. The PICS survey data is matched at the firm-level with the second source of data, the firms' tax reports submitted to the Mongolia tax office.

As a first shot at measuring the extent of underreporting by survey firms, we report the firms' responses to the question "what % of total sales the *typical establishment* in your area of activity reports for tax purposes" across city industry and firm size (Table 1). Firms report that the typical firms on average underreport 37.7% of their sales. Although underreporting in the city of Erdenet is significantly higher at 10% compared to Darkan and Hovd, the differences among the other 3 cities and across industry and firm size are not

TABLE 1
Mean % of underreporting in sales by the direct approach

	Variable	Obs.	Mean	Std. Dev.	Min	Max
City	*Ulaanbaatar*	179	37.7	29.2	0	95
	Darkhan	44	34.6	26.6	0	85
	Erdenet	46	43.5*	26.0	0	97
	Hovd	28	33.7	29.0	0	97
industry	*Manufacture*	146	37.9	29.4	0	97
	Construction	73	39.1	28.2	0	97
	Service	55	34.1	29.7	0	90
	Tourism	23	38.3	27.5	0	85
Firm size	*Small* (<10 workers)	91	37.4	30.6	0	97
	Medium (10-100 workers)	176	37.3	28.0	0	97
	Large (>100 workers)	29	41.5	29.5	0	90
total		297	37.7	28.8	0	97

Note: weighted by sampling weights;
* denotes significance at 10%*; Source: WB PICS Mongolia (2004)*

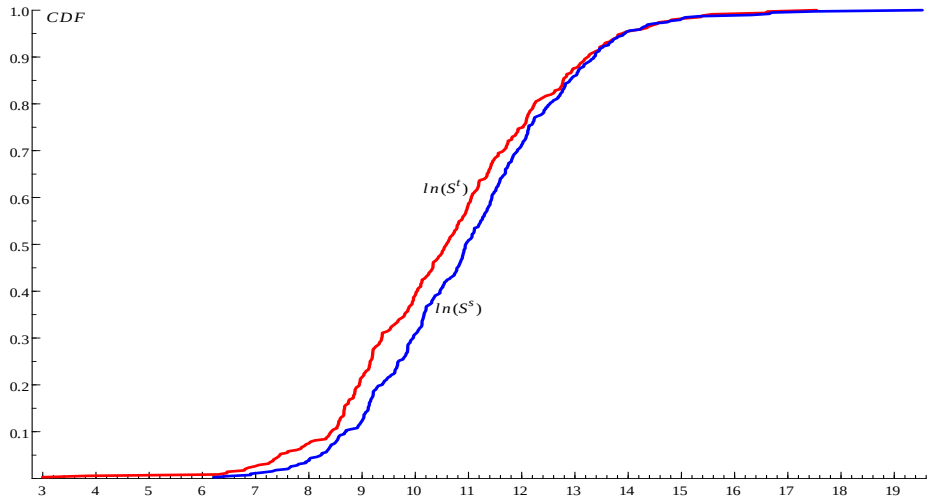
[22] See appendix A for more information on the data.

significant.[23] This finding confirms the conclusion of section 2 that the direct approach is less informative about the underreporting behavior of individual firms.

Our second shot at measuring underreporting is based on the so-called indirect approach by interpreting the discrepancy between the reported sales in the survey and the sales reported to the tax office as an indicator of underreporting. This approach can be seen as a special case of the MIMIC model, where one assumes that survey sales reflect true sales.[24]

Figure 2 shows the cumulative density functions for the reported sales to the tax office and in the survey. The distribution of sales reported in the survey first-order stochastically dominates the distribution of sales reported in the tax office. This implies that the total

Figure 2: Cumulative Densities of (log) Sales reported in Survey and to Tax Office



taxes actually paid to the tax office is less than the total tax that should have been paid, if taxes actually paid increase monotonously in sales reported to the tax office. Table 2

[23] Based on tests of equal means and allowing for heteroskedasticity.
[24] I.e. $d_2 = 0$ and $\beta_2 = 1$ in equation (11).

reports the extent of underreporting in sales to the tax office applying the indirect approach across cities, industries and firm sizes, with underreporting calculated as $100 \times \left(\frac{S^s - S^t}{S^s}\right)$. Estimated underreporting is much smaller than revealed by the direct approach (table 1) and averages about 15%. If survey sales reflect true sales and firms tend to underreport sales to tax office, the indicator should be between 0 and 100, but we observe negative values for 96 firms.[25] This suggests the presence of serious measurement error and/or underreporting in the survey sales.

TABLE 2
Average % of Sales Underreported by the Indirect Approach

		Obs.	Mean	Std. Dev.	Min	Max
City	*Ulaanbaatar*	139	14.9	41.5	-125.9	91.3
	Darkhan	50	18.2	55.3	-121.2	91.3
	Edernet	46	13.5	54.2	-154.3	90.1
	Hovd	20	16.4	49.8	-132.8	86.7
Industry	*Manufacture*	131	18.0	46.2	-121.2	91.3
	Construction	66	12.4	47.1	-154.3	91.3
	Service	45	19.9	36.5	-98.0	90.8
	Tourism	13	-2.1*	16.7	-22.4	64.5
Size	*Small*	69	25.7**	42.3	-121.2	90.6
	Medium	156	13.3	46.8	-154.3	91.3
	Large	29	4.4	28.3	-80.9	71.5
Total		255	14.8	45.3	-154.3	91.3

Note: the top/bottom 5% of the observations are not reported; all figures are weighted by sampling weights;
*, ** denote significance at 10% and 5% respectively; *Source: WB PICS Mongolia (2004) & Tax Office Data 2003*

The MIMIC approach has three main advantages over the direct and indirect approaches. First, the MIMIC approach uses more information by incorporating firm-level incentives to comply with taxes and a production function model linking true sales with observable firm-level production characteristics. Second, the MIMIC model allows for measurement errors in sales reported in the survey and to the tax office. And third, the MIMIC model allows for underreporting not only to the tax office but also in the survey.

[25] 37 out of the 97 negative indicators are below minus 10%.

A. Estimated MIMIC model

The MIMIC model eqs. (7), (10)-(12) include vectors X_1 and $X_2$to approximate for price and productivity differences across firms (lnμ) and for the unknown parameters P, δ, $\bar{\theta}$ and τ in the tax compliance model. We include the following variables in X_1 and X_2.

First, variables capturing firm size, location (city) and sector are included in both X_1 and X_2 because they may affect firm productivity and tax reporting behavior (Dabla-Norris, Gradstein and Inchuauste 2008).[26] Larger firms may be more productive than small firms and have more bargaining power to obtain more advantageous product and factor prices; smaller firms may underreport relatively more than large firms because the latter are more formal and subject to more public attention but they may also underreport less if they are more transparent; different cities have different regulation environments that not only affect firm productivity but also the reporting behaviors to the tax office by firms; firms are also expected to form different perceptions of local tax office efficiency especially across different sectors.

We also included proxies for average skill level, management experience and capacity utilization in X_1 and X_2 as possible determinants of productivity and tax reporting behavior. The skill level and management experience of a firm are included in X_2 as underreporting opportunities may vary with the skill intensity of production and managers with more experience may be more adept in reducing the tax burden when dealing with the tax office.

Finally we included proxies for the investment climate in X_1 and X_2 as investment constraints such as tax burden, corruption, regulation burden and credit constraints have

[26] See appendix A for exact variable definitions.

been shown to affect both firm performance and tax reporting behavior of firms (Pommerehne and Weck-Hannemann 1996, Johnson, et al. 2000, Dollar, Hallward-Driemeier and Mengistae 2005).[27] For this paper no suitable proxy for tax burden is available but we include a dummy for *corruption* with a value one (zero) implying that a firm has (not) paid bribes or provided informal gifts in 2003. Also a dummy variable *credit* is included equal to one if a firm faces a constraint to formal credit such as bank loans and overdraft. We also considered a list of variables approximating for the burden of regulation, such as the number of visits by agency inspectors, the time spent dealing with regulations by management, but they were insignificant in the empirical analysis and were omitted.

We are aware of the possible endogeneity problem with the IC variables. Firms may pay bribes to get away with underreporting and firms with larger official sales may have more access to finance than firms with smaller official sales. Dollar *et al.* (2005) have proposed taking the city-industry average of the IC variables to reduce the possible endogeneity.[28] We have used the same strategy, but found inflated estimates for the coefficients of the size and industry dummies due to severe multicollinearity problems. On the other hand, including the IC dummies directly in the regression does not affect the estimated parameters of the other variables much. This suggests that possible endogeneity bias, if any, is limited to the estimated coefficients of the IC dummies.

[27] The quality of government institutions or legal systems has been shown to also affect informality (Pommerehne and Weck-Hannemann 1996, Loayza 1996, Johnson, et al. 2000, Dabla-Norris, Gradstein and Inchuauste 2008).

[28] However as Dollar *et al*. (2005) also point out that even the city-sector averages won't be exogenous if more efficient firms self-select to better climate locations and the analysis does not fully control for all the forces driving the self-selection behaviors.

The next table reports the estimates for the MIMIC model using Maximum Likelihood estimation.[29] Panel A presents the results for the structural equation with dependent variable $\ln(S^*)$ (equation 7) and panel B and panel C show the results for the two measurement equations with dependent variables log of sales reported to the tax office (equation 10) and log of sales reported in the survey (equation 11) respectively. Panel D provides the multivariate normality tests on the residuals. Column (1) shows the results without IC variables, column (2) and column (3) add *credit* and *corruption* respectively, and column (4) adds both. Notice that we drop the constant d_1 in equation 10) as this allows for identification of the model without relying on the non-linear constraint $lsale_i \leq \ln(S_i^*)$ (see Appendix B) and, unlike the other two constants in the model, d_1 was found to be insignificant.

Panel A shows the estimates for the structural equation for the latent variable of true sales. The coefficients for the three industry dummies suggest that true sales in construction, service and tourism are respectively 32%, 26% and 171% higher than in manufacturing (column (4))[30], although only the coefficient for tourism is significant. The coefficients for *medium* and *large* are both significant at 5% level and imply that firms of medium and large sizes produce almost 0.6 and 2.7 times more than small firms respectively, holding everything else constant. Moreover, firms' sales increase with increasing capital and higher capacity utilization. The dummy variables *credit* and *corruption* control for the impact of the investment climate on productivity and prices and consequently on true sales. The

[29] The estimated coefficients of the city dummies in X_1 are small and (jointly) insignificant and were excluded from X_1 in the final specification. Similarly the estimated coefficients for Skill and Experience in X_1 and Capacity in X_2 were insignificant and small and therefore omitted from the final (more parsimonious) specification.

[30] Exp(0.279)-1=0.32, exp(0.234)-1=0.26, and exp(0.998)-1=1.71.

TABLE 3
Truncated MIMIC Model with Latent Variable $\ln(S^*)$

N = 231		LnL	-125.40	-121.94	-124.42	-121.17
			(1)	(2)	(3)	(4)
Panel A: $\ln(S^*)$		*Construction*	0.286	0.328*	0.231	0.279
			(0.181)	(0.178)	(0.187)	(0.184)
		Service	0.269	0.254	0.241	0.234
			(0.235)	(0.231)	(0.234)	(0.230)
		Tourism	0.840**	0.929**	0.925**	0.998**
			(0.335)	(0.342)	(0.371)	(0.373)
		Medium	0.584**	0.518**	0.577**	0.519**
			(0.228)	(0.227)	(0.226)	(0.224)
		Large	1.343**	1.245**	1.342**	1.254**
			(0.391)	(0.390)	(0.392)	(0.391)
		Capacity	0.271**	0.240**	0.266**	0.237**
			(0.118)	(0.116)	(0.118)	(0.116)
		Capital	0.330**	0.323**	0.332**	0.325**
			(0.038)	(0.037)	(0.038)	(0.037)
		Credit		-0.344**		-0.321**
				(0.155)		(0.158)
		Corruption			0.207	0.176
					(0.172)	(0.167)
		Constant	4.520**	4.768**	4.619**	4.845**
			(0.725)	(0.725)	(0.727)	(0.727)
Panel B: $\ln(S^t)$		**$\ln(S^*)$**	1	1	1	1
		Construction	0.921**	0.917**	1.046**	1.023**
			(0.381)	(0.369)	(0.407)	(0.389)
		Service	0.205	0.238	0.265	0.280
			(0.414)	(0.403)	(0.48)	(0.403)
		Tourism	-0.548	-0.557	-0.738	-0.705
			(0.619)	(0.626)	(0.711)	(0.699)
		Medium	0. 800*	0. 801**	0.806*	0.794**
			(0.426)	(0.414)	(0.423)	(0.407)
		Large	1.513**	1.519**	1.486*	1.480**
			(0.788)	(0.768)	(0.790)	(0.765)
		Darkan	-0.997**	-0.994**	-1.044**	-1.032**
			(0.366)	(0.355)	(0.369)	(0.355)
		Erdenet	-1.102**	-1.062**	-1.131**	-1.071**
			(0.416)	(0.396)	(0.421)	(0.397)
		Hovd	-0.955**	-0.853*	-1.005**	-0.896**
			(0.486)	(0.461)	(0.487)	(0.461)
		Skill	0.750**	0.729**	0.726**	0.706**
			(0.243)	(0.233)	(0.241)	(0.231)
		Experience	-0.270**	-0.262**	-0.280**	-0.269**
			(0.112)	(0.107)	(0.112)	(0.107)
		Credit		-0.020		-0.058
				(0.271)		(0.277)
		Corruption			-0.359	-0.315
					(0.317)	(0.298)
Panel C: $\ln(S^s)$		**$\ln(S^*)$**	0.665**	0.649**	0.680**	0.663**
			(0.066)	(0.067)	(0.066)	(0.067)
		Constant	1.666**	1.809**	1.646**	1.798**
			(0.612)	(0.624)	(0.592)	(0.608)
Panel D: *Multivariate*		*Chi-square test (4)*	5.242	5.085	4.979	5.033
Normality test		*p-value*	0.263	0.279	0.289	0.284
Asymptotic		*Chi-square test (4)*	5.025	4.982	4.477	4.940
Multivariate		*p-value*	0.285	0.289	0.345	0.294
Normality test						

Note: dependent variables are **$\ln(S^*)$**, **$\ln(S^t)$** & **$\ln(S^s)$**, standard errors in the brackets , * and** denote significant at 10% and 5%,; *Source: same as Table 2*

impact of *credit* is negative and significant, implying that firms with a credit constraint produce 27.5% less than firms without the constraint. The proxy for corruption burden has an unexpected positive sign but is small and insignificant.

Panel B and C show the results of the two measurement equations with indicators $\ln(S^t)$ and $\ln(S^s)$. In the absence of underreporting, the coefficients in panel B (except for true sales) should be zero, while in Panel C the coefficient of $\ln(S^*)$ should be equal to one and the constant should be zero. But most coefficients in panel B and C are significantly different from zero (different from one for $\ln(S^*)$ in Panel C) and have the expected signs. Therefore underreporting in sales exists and also varies across firms.

All the parameter estimates in panel B for the impacts of variables on sales reported to the tax office are obtained conditional on true sales ($\ln S^*$), and the marginal effects of the variables should be interpreted accordingly. Firms in construction report significantly more to the tax office than firms in the manufacturing sector, ceteris paribus. Also medium and large firms report significantly more, everything else equal and conditional on true sales. The parameter estimates for Darkan, Erdenet and Hovd are all negative and significant and hence firms outside of Ulaanbaatar report significantly less sales to the tax office conditional on true sales. Moreover, firms with a higher average skill level also report significantly larger sales compared to other firms, while firms with more experienced managers tend to report less sales to the tax office. Both investment climate dummies appear small and insignificant and hence we do not have evidence that credit constraints and corruption burden affect underreporting to the tax office at given levels of true sales.

Panel C shows the results for the measurement equation using the estimated sales from the survey as indicator.[31] The coefficient of the log of true sales is 0.66 and significantly different from zero and one at the 5% significance level. Therefore survey sales are significantly and positively correlated with true sales, but with a correlation smaller than one. This does not imply underreporting, however, as the constant term is also positive and significant at the 5% level, and for small values of true sales this would suggest over-reporting. However, based on the estimated conditional predictions of the firms' true sales (see next subsection), most of the firms (74%) are estimated to underreport sales in the survey.

The above estimation of the MIMIC model assumes a multivariate normal distribution for the disturbances. When this assumption is violated, the estimated parameter values from truncated maximum likelihood become inconsistent (Cameron and Trivedi 2005). We test the multivariate normality assumption on the reduced-form MIMIC model residuals with the results shown in panel D. Both the multivariate and asymptotic multivariate normality tests do not reject the null hypothesis of multivariate normality. We have also estimated the MIMIC model with different choices for the capital return and depreciation rates used in the calculation of the capital variable (equation 6), and we find that our results are robust with respect to plausible alternative parameter choices.

[31] We've tried specifications including various firm characteristics in panel C similar as Panel B, but the coefficients all appear to be very small and insignificant.

B. Prediction of True Sales

The main objective of this paper is to estimate a firm's true sales and therefore the extent of underreporting in sales to the tax office and in the survey. Prediction of the latent variable true sales in the MIMIC model can be done conditional on either the causal variables X only or on both X and the indicators.[32] Previous MIMIC empirical literature predicts the latent variable(s) conditional on the causal variables *X* only (e.g., Giles 1999a, Chaudhuri *et al*. 2006, Dell'Anno *et al*. 2007). In order to increase precision, we choose to predict conditionally on both the causal variables *X* and the indicators $ln(S^t)$ and $ln(S^s)$and calculate $\mathrm{E}\left(\mathrm{S}^*|\widetilde{\mathbf{X}}, \ln(\mathrm{S}^{\mathrm{t}}), \ln(\mathrm{S}^{\mathrm{s}}), \ln(\mathrm{S}^{\mathrm{t}}) \leq \ln(\mathrm{S}^*)\right)$.[33] Next we calculate the unconditional underreporting to the tax office and in the survey by each firm in our sample using the formula $100 \times \frac{\mathrm{E}\left(\mathrm{S}^*|\widetilde{\mathbf{X}}, \ln(\mathrm{S}^{\mathrm{t}}), \ln(\mathrm{S}^{\mathrm{s}}), \ln(\mathrm{S}^{\mathrm{t}}) \leq \ln(\mathrm{S}^*)\right) - \mathrm{S}^{\mathrm{i}}}{\mathrm{E}\left(\mathrm{S}^*|\widetilde{\mathbf{X}}, \ln(\mathrm{S}^{\mathrm{t}}), \ln(\mathrm{S}^{\mathrm{s}}), \ln(\mathrm{S}^{\mathrm{t}}) \leq \ln(\mathrm{S}^*)\right)}$, $\mathrm{i} = \mathrm{t}, \mathrm{s}$. The results are reported in Table 4.

Underreporting to the tax office is on average larger than in the survey (38.6% versus 11.9%). With respect to underreporting to the tax office, firms that are small (42.8%), firms that are located in Darkhan or Edernet (46.6% and 46.2%), and firms that do pay bribes and/or provided informal gifts when dealing with government authorities (41.9%), tend to underreport significantly more than other firms. Also, and surprisingly, firms in the construction sector tend to underreport less (33.9%). Firms facing credit constraints underreport slightly more on average than firms without such constraint, but the difference is not significant (39.9% v. s 37.8%).

[32] Joreskog and Goldberger (1975) derived the formula of prediction conditional on both causal variables and indicators. But since we impose an inequality constraint, their formula for the conditional expectation will have to be modified. Another (minor) difference is that we normalize by taking $\beta_1 = 1$ instead of $\mathrm{var}(\mathrm{u}) = 1$.

[33] For the exact derivation of this conditional expectation see appendix C.

We also observe underreporting in the survey albeit at a lower level. Underreporting in the survey is especially high for firms in Darkhan and Edernet (23.2% and 24.8%), in the tourism sector (37.6%), and for large firms (32.7%). The high level of underreporting for

TABLE 4
Mean % of Total Sales Underreported by MIMIC Approach

	Underreporting in Sales		**To tax office**		**In the survey**	
		Obs.	Mean	Std. Dev.	Mean	Std. Dev.
City	*Ulaanbaatar*	126	36.9	22.4	9.4	38.0
	Darkhan	50	46.6**	24.7	23.2**	29.9
	Edernet	38	46.2**	22.4	24.8**	29.2
	Hovd	17	38.0	20.5	9.5	38.6
Industry	*Manufacture*	125	42.4	24.0	13.3	32.0
	Construction	62	33.9**	18.4	5.3*	44.4
	Service	30	36.7	21.7	14.1	26.3
	Tourism	14	43.7	24.2	37.6**	24.8
Size	*Small*	57	42.8	24.4	0.9	30.7
	Medium	152	37.4*	21.8	12.6**	39.3
	Large	22	37.6	16.4	32.7**	19.9
Corruption payment	*no*	121	35.3	20.2	8.7	39.8
	yes	110	41.9**	23.3	15.2*	33.8
Credit constrained	*no*	151	37.8	21.3	17.0**	35.3
	yes	80	39.9	23.4	2.8	38.4
Total		231	38.6	22.0	11.9	37.0

Note: * and ** denote significance of two-sample t-tests for equal means with different observations and variances at 10% and 5% level respectively; figures are weighted by sampling weights; *Source*: *calculated using MIMIC estimation results of Table 3 (column (4))*

firms in Darkhan and Edernet may reflect the perceived lower quality of the survey team in these locations.[34] The high level of underreporting in the survey by large firms may be explained by the fact that the estimated sales are based on the three main products (see Appendix A) while large firms are more likely to have more than 3 product lines. Underreporting is also relatively high for firms that pay bribes and/or provided informal gifts when dealing with government authorities (15.2%) and for firms that are credit constrained (17.0%).

Table 4 shows the bivariate relationship between underreporting behavior and firm characteristics (and investment climate constraints). It is therefore useful to investigate

[34] Based on personal communication with supervisor of the survey.

which of the factors affect underreporting most. Firm size and location are correlated and investment climate variables are also correlated with firm characteristics. Therefore we also did a descriptive regression analysis with underreporting to the tax office as the dependent variable and firm characteristics as well as the IC constraint variables as independent variables, allowing for heteroskedasticity.[35] The resulting regression shows that the size effect of underreporting disappears with the inclusion of other variables. Firms with more capital underreport significantly less to the tax office while corruption and credit constraints increase underreporting to the tax office. Also firms outside Ulaanbaatar tend to underreport more but less if they are active in the construction sector.

Finally we calculated the percent of *aggregate* sales underreported to the tax office and in the survey. Aggregate underreporting is 37.5% to the tax office and 22.8% in the survey. These figures are respectively lower and higher than the mean firm-level underreporting reported in Table 4, because underreporting decreases in firms size for sales reported to the tax office but increases for sales reported in the survey.

5 Discussion

We have used matched firm-level survey and official tax data to estimate the true sales and the extent of underreporting in sales by formal firms in Mongolia. Three different approaches have been explored, namely a direct approach, an indirect approach, and a

[35] Underreporting to tax office (as proportion of predicted true sales) $=$ $.01(.04)\text{medium} - .01(.06)\text{large} + .08^{**}(.04)\text{Darkan} + .05(.04)\text{Erdenet} + .02(.06)\text{Hovd} - .12^{**}(.03)\text{construction} - .06(.05)\text{service} + .03(.07)\text{tourism} - .01^{*}(.01)\text{capital} - .01(.02)\text{skill} + .06^{*}(.03)\text{credit} + .09^{**}(.03)\text{corruption}$ (robust standard errors in the brackets).

modeling (MIMIC) approach. These approaches have been applied widely in the shadow economy literature but primarily for macro data (for the indirect and modeling approach) and without a proper economic foundation for the MIMIC approach.

We argue that our MIMIC approach provides the more accurate estimate of the extent of underreporting because it incorporates firm-level incentives to comply with taxes and a production function technology linking true sales with observable firm-level production characteristics. It also controls for measurement errors and allows for underreporting in both official tax and survey data.

The next table compares the results from the three approaches. For the direct approach, the predicted true sales is calculated by dividing the sales reported to the tax-office by the percentage of sales that are reported to the tax office by a typical firm according to the same firm (multiplied by 100). For the indirect approach, the predicted true is assumed to be equal to the survey sales and for the MIMIC approach it is equal to the conditional expectation. Comparing the predicted true sales with the actual sales reported to the tax office, we have a total of 186 firms for which we have an estimate of underreporting from each of the approaches (Table 5).

The indirect approach gives the smallest estimates of underreporting to the tax office. This low estimate reflects the sensitivity of the indirect approach to measurement error.[36] It

[36] Measurement error and convexity in survey sales implies that $E\left(\frac{sale\_t}{sale\_s}\right) > \frac{sale\_t}{E(sale\_s)}$, and therefore the expected level of underreporting ($E\left(1 - \left(\frac{sale_t}{sale_s}\right)\right) \times 100$) is biased downward.

also suffers from underreporting in the survey (Table 4). The direct approach gives comparable but somewhat lower estimates than the MIMIC approach for mean underreporting at the firm-level, while for aggregate underreporting the estimate is clearly higher. The relative high estimate for aggregate underreporting by the direct approach

TABLE 5
Comparison of Three Approaches to Measuring Underreporting

	Firm-level (mean)	Aggregate
Direct Approach	35.1%	43.9%
Indirect Approach	15.2%	17.4%
MIMIC Approach	38.0%	36.9%

Note: all figures are weighted by sampling weights

follows from the fact that the predicted true sales in the direct approach are calculated by dividing the reported tax office sales of a firm by the percentage of sales the "typical establishment" reports to the tax office according to the same firm (multiplied by 100). Consequently a higher reported percentage of underreporting gives a higher estimate of true sales for the same firm and therefore is weighted more in the calculation of aggregate underreporting.

In sum, the MIMIC approach appears to give the more accurate estimate of underreporting. The indirect approach performs poorly and underestimates underreporting because of being sensitive to measurement errors and underreporting in the survey data. The direct approach gives an estimate of the firm-level average percentage of underreporting which is somewhat lower than the MIMIC approach, confirming the conjecture of Schneider and Enste (2000, p.92) that the direct approach provides a lower bound for true underreporting. However, the direct approach gives a too high estimate for aggregate

underreporting because of measurement error and appears less useful as an indicator of underreporting by individual firms (section 2).

The MIMIC model approach also shows that underreporting is systematically related to firm characteristics and investment climate variables. This is true for underreporting to the tax office but also for sales reported in the survey. Even though our analysis is based on cross sectional data and we are not able to fully control for firm heterogeneity, our finding that firms also underreport in the survey may pose a serious challenge for economic analyses that use firm-level survey data without adequately controlling for possible and systematic underreporting bias.

References

Aigner, Dennis, C.A. Knox Lovell, and Peter Schmidt. "Formulation and Estimation of Stochastic Frontier Production Function Models." *Journal of Econometrics*, 1977: Vol. 6, No. 1, 21-37.

Allingham, Michael, and Agnar Sandmo. "Income Tax Evasion: a Theoretical Analysis." *Journal of Public Economics*, 1972: Vol. 1, 323-338.

Anderson, James H. "The Size, Origins, and Character of Mongolia's Informal Sector During the Transition." *World Bank Policy Research Working Paper No. 1916*, 1998: 1-64.

Bartelsman, Eric J., and Mark E. Doms. "Understanding Productivity: Lessons from Longitudinal Microdata." *Journal of Economic Literature*, 2000: Vol. 38, No. 3, 569-594.

Bigsten, Arne, et al. "Credit Constraints in Manufacturing Enterprises in Africa." *Journal of African Economics*, 2003: Vol. 12, No. 1, 104-125.

Bikales, B., C. Khurelbaatar, and K. Schelzig. *The Mongolian Informal Sector: Survey Results and Analysis.* Ulaanbaatar: Economic Policy Project, DAI, 2000.

Cameron, A. Collin, and Pravin K. Trivedi. *Microeconometrics: methods and applications, sections 16.2 - 16.3, pp. 530-544.* New York: Cambridge University Press, 2005.

Chaudhuri, Kausik, Friedrich Schneider, and Sumana Chattopadhyay. "The Size and Development f the Shadow Economy: An Empirical Investigation from States of India." *Journal of Development Economics*, 2006: Vol. 80, 428-443.

Chen, Kong-pin, and C.Y. Cyrus Chu. "Internal Control versus External Manipulation: a Model of Corporate Income Tax Evasion." *The Rand Journal of Economics*, 2005: Vol. 36, No. 1, 151-164.

Christensen, Laurits R., and Dale W. Jorgensen. "The Measurement of U.S. Real Capital Input, 1929-1967." *Review of Income and Wealth*, 1969: Vol. 15, 293-320.

Crocker, Keith J., and Joel Slemrod. "Corporate Tax Evasion with Agency Costs." *Journal of Public Economics*, 2005: vOL. 89, No. 9-10, 1593-1610.

Cull, Robert, and Lixin Colin Xu. "Institutions, ownership, and Finance-the determinants of profit reinvestment among Chinese firms." *Journal of Financial Economics*, 2005: Vo. 77, 117-146.

Dabla-Norris, Era, and Junko Koeda. "Informality and Bank Credit: Evidence from Firm-Level Data." *IMF Working Paper*, 2008: 1-37.

Dabla-Norris, Era, Mark Gradstein, and Gabriela Inchuauste. "What Causes Firms to Hide Ouput? The Determinants of Informality." *Journal of Development Economics*, 2008: vol. 85, 1-27.

Dell'Anno, Roberto, Miguel Gomez-Antonio, and Angel Pardo. "The Shadow Economy in Three Mediterranean Countries: France, Spain, and Greece. A MIMIC Approach." *Empirical Economics*, 2007: Vol. 33, 51-84.

Dollar, David, Mary Hallward-Driemeier, and Taye Mengistae. "Investment Climate and Firm Performance in Developing Economics." *Economic Development and Cultural Change*, 2005: Vol. 54, No. 1, 1-31.

Erard, Brian, and Jonathan S. Feinstein. "Hoesty and Evasion in the Tax Compliance Game." *Rand Journal of Economics*, 1994a: Vol. 25, No. 1, 1-19.

Frey, Bruno S., and Hannelore Weck-Hannemann. "The Hidden Economy as an 'Unobserved' Variable." *European Economic Review*, 1984: Vol. 26, 33-53.

Gatti, Roberta, and Maddalena Honorati. "Informality among Formal Firms: Firm-level, Cross-country Evidence on Tax Compliance and Access to Credit." *World Bank Policy Research Working Paper No. 4476*, 2008: 1-34.

Giles, David E.A. "Measuring the Hidden Economy: Implications for Econometric Modeling." *The Empirical Journal*, 1999: Vol. 109, 370-380.

Graetz, Michael J., and Jennifer F., Wilde, Louis L. Reinganum. "The Tax Compliance Game: Toward an Interactive Theory of Tax Enforcement." *Journal of Law Economics Organization*, 1986: Vol. 2, No.1, 1-32.

Grasmick, Harold, and Robert J. Bursick. "Conscience, Significant Others, and Rational Choice: Extending the Deterrence Model." *Law and Society Review*, 1990: Vol. 24, No. 3, 837-861.

Johnson, Simon, Daniel Kaufmann, John McMillan, and Christoper Woodruff. "Why do Firms Hide? Bribes and Unofficial Activity after Communism." *Journal of Public Economis*, 2000: Vol. 76, 495-520.

Joreskog, Karl G., and Arthur S. Goldberger. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association*, 1975: Vol. 70, No. 351, 631-639.

Loayza, Norman V. "The Economics of the Informal Sector: a Simple Model and Some Empirical Evidence from Latin America." *Carbegue-Rochester Conference Series on Public Policy 45.* North-Holland, 1996. 129-162.

Pommerehne, Werner W., and Hannelore Weck-Hannemann. "Tax Rates, Tax Administration and Income Tax Evasion in Switzerland." *Public Choice*, 1996: Vol. 88, 161-170.

Reinganum, Jennifer F., and Louis L. Wilde. "Income Tax Compliance in a Principle-Agent Framework." *Journal of Public Economics*, 1985: Vol. 26, No. 1, 1-18.

Schneider, Friedrich, and Dominik H. Enste. "Shadow Economies: Size, Causes, and Consequences." *Journal of Economic Literature*, 2000: vol. 38, 77-114.

Siegel, Donald. "The Impact of Computers on Manufacturing Productivity Growth: A Multiple-Indicators, Multiple-Causes Approach." *The Review of Economics and Statistics*, 1997: Vol. 79, No. 1, 68-78.

Teal, Francis. "The Size and Sources of Economic Rents in a Developing Country Manufacturing Labor Market." *The Economic Journal*, 1996: Vol. 106, 963-976.

Van de Ven, Wynand P.M. M., and Jacques Van der Graag. "Health as an Unobservable: a MIMIC Model of Demand for Health Care." *Journal of Heatlh Economics*, 1982: Vol. 1, No. 2, 157-183.

Appendix A: PICS Sample, Variable Definitions and Summary Statistics

PICS Sample

Initially 562 firms were randomly sampled from the business register. A few sampled firms could not be surveyed for a number of reasons, leaving 400 valid firms in the initial sample. [37] Out of this sample, interviews with 287 firms were completed successfully. [38] Based on the information provided by the tax office on stably operating establishments, 106 more firms were identified and interviewed, leaving 393 firms in the final sample. Post-stratification sampling weights were constructed to control for 1) varying sampling probabilities across strata defined by sector and location, 2) inaccuracies in the business register because firms either could not be found, were out of business, did change activities, or moved to different locations, and 3) non-response. The post stratification sampling weights for the 321 firms that were in the business register are equal to the proportion of the listed firms in the business register over the valid sampling frame by sector and city and the sampling weight was set equal to one for each of the other 72 firms that were not in the business register due to the inaccuracy of the frame.

The survey covers the three biggest cities located in the central and northwestern areas of Mongolia (Ulaanbaatar, Darkan and Erdenet) and 2 secondary cities in the West and East (Choibalsan and Hovd). During the survey process, the survey data collected from Choibalsan turned out to be poor and was replaced by information from original financial statements provided by the firms. This makes the survey information from Choibalsan less comparable with the information from the other cities, because the original financial statements are often used for tax purposes and an independent survey measure of sales of firms from Choibalsan is missing. Hence firms from Choibalsan were excluded from our analysis.

Variable Definition and Construction

1) **S^s ($\ln(S^s)$)**: (log of) total sales for 2003 from PICS Mongolia. Data on total sales is not directly available in the PICS survey data. However firms were asked to report the sales

[37] They are omitted because they couldn't be found, out of business, changed activity to outside the selected sector, or moved to a different region.

[38] Not all sampled firms were successfully interviewed due to several reasons: 1) they cooperated but provided poor data, 2) they were contacted but no interview could be done (due to elections, holidays, vocation, et cetera), 3) they could not be found, 4) they were out of business, 5) they refused to cooperate, 6) they changed activity and were operating outside selected sectors, 7) for other reasons.

value of the three most important products/services. They were also asked about the importance of these three products/services as a percentage of total sales. [39] Total sales is calculated as:

$$S^s = 100 \times \frac{\sum_{i=1}^{3} \text{sales of the } i_{th} \text{ important product in 2003}}{\sum_{i=1}^{3} \% \text{ of importance of the sales of product } \boldsymbol{i} \text{ in total sales in 2003}}$$

Firms from the sample which reported only one product and for which the tax office sales equal the survey sales were excluded from the sample as these firms were most likely to give the same sales data in the survey as reported to the tax office. [40]

2) **S^t ($\ln(S^t)$)**: (log of) total sales reported to the Mongolian tax office for the year 2003 by Mongolian firms. The reported total sales are the sum of the reported sales from firm's main activity and sales from other activities in 2003.
3) ***skill:*** log of mean wage (total labor costs from the survey divided by total employment size), measuring a firm's average skill level of workers.
4) ***capital***: log of capital service flow, calculated from gross book values of the capital (buildings, machinery and equipment) with straight line depreciation rates (2.5% for buildings and 10% for machinery and equipments) plus the risk free bond rate (14%) in Mongolia (equation 6).
5) ***experience:*** manager's experience in (logarithm) years.
6) ***capacity:*** firms' capacity utilization.
7) ***credit:*** a dummy variable with 1 indicating constraint in access to finance and 0 indicating no constraint in access to finance. *Credit* is equal to zero when a firm received a term loan in 2003 or when a firm did not apply for a term loan from a bank or financial institution because it did not need term loans; otherwise it is equal to one when no term loan was received because an application was turned down.
8) ***corruption****:* a dummy variable with 1 indicating firm has ever paid bribes/informal gifts in the year 2003 for receiving service or approval, obtaining licenses, dealing with regulation agencies, securing government contract, or getting import/export customs clearance, et cetera.
9) *Other dummies*: 4 **city** dummies (*Ulaanbaatar*, *Darkan*, *Erdenet* and *Hovd*) indicating locations; 4 **industry** dummies (*manufacture, construction, service*, and *tourism*) indicating sectors; 3 **firm size** dummies defined by employment size (*small* (<10), *medium* (10-100), and *large* (>100)) referring to small, medium and large firms when the dummy equals one respectively. As over a third of the firms hire seasonal workers

[39] Firms were also asked to provide the raw material costs of the three most important inputs and the corresponding % of importance in total input costs. Firms were not asked about their total sales from the balance and income sheet because they were less willing to provide this information.

[40] Excluding these firms do not really affect the results.

with fulltime equivalent working days ranging from 20 to 312 days, we calculate a firm's employment size by the sum of the number of permanent workers and the weighted number of temporary workers.[41]

Summary Statistics

Table A.1 summarizes the data used in the MIMIC model. The majority of the firms are of small or medium sizes. Moreover, over half of the firms are located in Ulaanbaatar with the rest located in the other three cities, and most firms concentrate in the manufacturing and construction industries.

TABLE A.1 Summary Statistics for the MIMIC Model

Variable	Mean	Std. Dev.	Min	Max
log($\mathbf{S^s}$)	11.11	1.40	8.33	13.97
log($\mathbf{S^t}$)	10.68	1.58	7.45	14.07
Capital	8.77	2.01	3.65	14.60
Skill	6.52	.71	2.77	9.02
Experience	1.64	1.32	0	3.71
Capacity	4.13	.56	1.10	4.62
City (dummies)				
Ulaanbaatar	.55	.50	0	1
Darkan	.22	.41	0	1
Erdenet	.16	.37	0	1
Hovd	.07	.26	0	1
Size (dummies)				
Small	.25	.43	0	1
Medium	.66	.48	0	1
Large	.10	.29	0	1
Sector (dummies)				
Manufacture	.54	.50	0	1
Construction	.27	.44	0	1
Service	.13	.34	0	1
Tourism	.06	.24	0	1
IC (dummies)				
Corruption	.48	.50	0	1
Credit	.35	.48	0	1
Sample size	231			

[41] The weight used is the ratio of total number of days worked per worker (in full-time equivalent days) for temporary workers over the total number of days worked per worker for permanent workers in 2003.

Appendix B: MIMIC Model Identification and Estimation by Maximum Likelihood

The MIMIC model is given by equations (7), (10)-(12). For notational convenience, we put the non-overlapping variables in X_1 and X_2 into a new $k \times 1$ vector denoted by X and the corresponding parameter vectors b and δ_2 are written as b^n and δ_2^n.[42] Plugging the structural equation (7) into the two measurement equations (10) and (11), and rewriting the nonlinear constraint (12) using (10), we derive the following reduced form model with truncation:

$$y_1 \equiv \ln S^r = a + d_1 + X(b^n + \delta_2^n) + c \ln K_v^* + \varepsilon_t + u \qquad \text{if } d = 1 \tag{13}$$

$$y_2 \equiv \ln S^c = \beta_2 a + d_2 + X b^n \beta_2 + \beta_2 c \ln K_v^* + \varepsilon_s + \beta_2 u \qquad \text{if } d = 1 \tag{14}$$

$$y_3^* \equiv \varepsilon_t + d_1 + X\delta_2^n \tag{15}$$

$$d = 1(y_3^* < 0) \tag{16}$$

The nonlinear constraint is transformed to $\varepsilon_t \leq -d_1 - X\delta_2^n$ and ε_t is truncated (from above) normal. Without the nonlinear constraint $\varepsilon_t \leq -d_1 - X\delta_2^n$, the model is not fully identified. We can identify β_2 and c from the natural logarithm of capital service flow ($\ln K_v^*$) and this allows us to identify b^n from $b^n\beta_2$ and then δ_2^n from $(b^n + \delta_2^n)$. But it is still impossible to identify a, d_1 and d_2 separately. The identification problem is solved when the nonlinear constraint kicks in, through which the parameter d_1 is identified, then a can be identified from $a + d_1$, and finally d_2 is identified from $\beta_2 a + d_2$. Therefore the complete identification of the model crucially relies on the nonlinear constraint. However, identification is possible without the nonlinear constraint if we can drop one of the three constant terms (a, d_1, d_2).

Rewriting the reduced form model in matrix form gives:

$$\begin{aligned}(y_1 \quad y_2) &= (1 \quad X \quad \ln K_v^*)\begin{pmatrix} a + d_1 & \beta_2 a + d_2 \\ b^n + \delta_2^n & b^n\beta_2 \\ c & \beta_2 c \end{pmatrix} + (u + \varepsilon_t \quad \beta_2 u + \varepsilon_s) \\ &= \widetilde{\mathbf{X}}(\boldsymbol{\beta} \quad \boldsymbol{\gamma}) + (u + \varepsilon_t \quad \beta_2 u + \varepsilon_s) \qquad \text{if } d = 1 \\ y_3^* &= \varepsilon_t + d_1 + X\delta_2^n, \qquad d = 1(y_3^* < 0)\end{aligned}$$

We denote the variance-covariance matrix of the three disturbances $(\varepsilon_t, \varepsilon_t + u, \varepsilon_s + \beta_2 u)$ by $\boldsymbol{\Sigma}$. Taking into account the truncation condition $d = 1(y_3^* < 0)$, ε_t will be truncated normal. If u is uncorrelated with ε_t, then the disturbance $u + \varepsilon_t$ consists of a normally

[42] If $X_1 \neq X_2 \neq X$, then $b^n \neq b$ and $\delta_2^n \neq \delta_2$, for variables that do not appear in X_1 but in X, the corresponding coefficients in b^n will be zero, the rest of the elements in b^n are the same as in b; and it's similar for δ_2^n.

distributed error term u and a truncated normally distributed ε_t. In this case the structural equation fits within the stochastic frontier production function approach proposed by Aigner, Lovell and Schmidt (1977), where deviations from the production function frontier derive from two sources, namely a term truncated at zero (half-normal) approximating for inefficiency and a firm-specific idiosyncratic shock from a normal distribution with mean zero. In our case the truncated term ε_t does not reflect inefficiency, however, but the fact that firms have an incentive to underreport. This suggests that existing estimates of inefficiency based on the stochastic frontier production function approach may reflect underreporting behavior rather than genuine technical inefficiencies to some extent. In practice our model is more complicated than the stochastic frontier production function approach, with a truncation threshold depending on unobserved parameters and varying across firms, and allowance for correlation between u and ε_t.

The model can be estimated by maximum likelihood. First of all, the probability that y_{1i} and y_{2i} are observed for firm i is given by:

$$\Pr\left(y_{1i}, y_{2i} | y_{3i}^* < 0, \widetilde{\mathbf{X}}\right) = \frac{\Pr\left(y_{1i}, y_{2i}, y_{3i}^* < 0 \middle| \widetilde{\mathbf{X}}\right)}{\Pr\left(y_{3i}^* < 0 \middle| \widetilde{\mathbf{X}}\right)} = \frac{f\left(y_{1i}, y_{2i} \middle| \widetilde{\mathbf{X}}\right) \Pr\left(y_{3i}^* < 0 \middle| y_{1i}, y_{2i}, \widetilde{\mathbf{X}}\right)}{\Pr\left(y_{3i}^* < 0 \middle| \widetilde{\mathbf{X}}\right)}$$
$$= \frac{f(u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si}) \Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n | u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})}{\Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n)}$$

Taking the logarithmic transformation of the probability and summing the log likelihood over all firms in the sample, we obtain the log likelihood:

$$\ln L = \sum_{i=1}^{N} \Big(\ln\big(f(u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})\big) + \ln(\Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n | u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})) - \ln(\Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n)) \Big)$$

$$= \sum_{i=1}^{N} \left(\ln\big(f(u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})\big) + \ln\left(\int_{-\infty}^{-d_1 - X_i\delta_2^n} f(\varepsilon_{ti} | u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si}) d\varepsilon_{ti} \right) - \ln\left(\int_{-\infty}^{-d_1 - X_i\delta_2^n} f(\varepsilon_{ti}) d\varepsilon_{ti} \right) \right)$$

$\Pr\left(y_{3i}^* < 0 \middle| \widetilde{\mathbf{X}}\right) = \Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n)$ is the probability of truncation for firm i. Assume that the three disturbances $\varepsilon_s, \varepsilon_t$ and u are multivariate normally distributed with mean zero. Therefore $(\varepsilon_t, u + \varepsilon_t, \beta_2 u + \varepsilon_s)$ is also multivariate normally distributed with mean zero and variance-covariance matrix denoted by $\boldsymbol{\Sigma}$ and $(\varepsilon_t | u + \varepsilon_t, \beta_2 u + \varepsilon_s)$ is also normally

distributed. Let $\mathbf{v^T} = \begin{pmatrix} u + \varepsilon_t \\ \beta_2 u + \varepsilon_s \end{pmatrix}$ with variance-covariance matrix $\mathbf{\Sigma_{22}}$ and the conditional density is $f(\varepsilon_{ti}|u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si}) = f(\varepsilon_{ti}|\mathbf{v}_i)$. Based on the multivariate normality assumption, the conditional mean $\mu_{\varepsilon_t|\mathbf{v}}$ and variance $\sigma^2_{\varepsilon_t|\mathbf{v}}$ of $(\varepsilon_t|\mathbf{v})$ are derived to be $\Sigma_{12}\Sigma_{22}^{-1}\begin{pmatrix} y_1 - \widetilde{\mathbf{X}}\boldsymbol{\beta} \\ y_2 - \widetilde{\mathbf{X}}\boldsymbol{\gamma} \end{pmatrix}$ and $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ respectively, where $\text{var}(\varepsilon_t) = \Sigma_{11}, \text{cov}(\varepsilon_t, \mathbf{v}) = \Sigma_{12}$. The probability of the i_{th} observation is:

$$\Pr(y_{1i}, y_{2i}|y^*_{3i} < 0, X) = \frac{\Pr(y_{1i}, y_{2i}, y^*_{3i} < 0|\widetilde{\mathbf{X}})}{\Pr(y^*_{3i} < 0|\widetilde{\mathbf{X}})} = \frac{f(y_{1i}, y_{2i}|\widetilde{\mathbf{X}})\Pr(y^*_{3i} < 0|y_{1i}, y_{2i}, \widetilde{\mathbf{X}})}{\Pr(y^*_{3i} < 0|\widetilde{\mathbf{X}})}$$
$$= \frac{f(u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})\Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n|u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})}{\Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n)}$$

Based on the multivariate normality assumption, we have:

a) $\Pr(\varepsilon_{ti} < -d_1 - X_i\delta_2^n|u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si}) = \int_{-\infty}^{-d_1 - X_i\delta_2^n} f(\varepsilon_{ti}|u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si})d\varepsilon_{ti} = \int_{-\infty}^{-d_1 - X_i\delta_2^n} \phi\left(\frac{\varepsilon_{ti} - \mu_{\varepsilon_t|v}}{\sigma_{\varepsilon_t|v}}\right)d\varepsilon_{ti} = \Phi\left(\frac{-d_1 - X_i\delta_2^n - \mu_{\varepsilon_t|v}}{\sigma_{\varepsilon_t|v}}\right)$

b) $f(u_i + \varepsilon_{ti}, \beta_2 u_i + \varepsilon_{si}) = \frac{1}{\sqrt{2\pi|\Sigma_{22}|}}\exp\left(-\frac{1}{2}\begin{pmatrix} y_{1i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta} \\ y_{2i} - \widetilde{\mathbf{X}}_i\boldsymbol{\gamma} \end{pmatrix}^T {\Sigma_{22}}^{-1}\begin{pmatrix} y_{1i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta} \\ y_{2i} - \widetilde{\mathbf{X}}_i\boldsymbol{\gamma} \end{pmatrix}\right)$

c) $\Pr(\varepsilon_{ti} < -X_i\delta_2^n) = \int_{-\infty}^{-d_1 - X_i\delta_2^n} f(\varepsilon_{ti})d\varepsilon_{ti} = \int_{-\infty}^{-d_1 - X_i\delta_2^n} f\left(\frac{\varepsilon_{ti}}{\sqrt{\Sigma_{11}}}\right)d\varepsilon_{ti} = \Phi\left(\frac{-d_1 - X_i\delta_2^n}{\sqrt{\Sigma_{11}}}\right)$

Plugging equations a)-c) into the probability function, taking the logarithm of the probability, and summing over firms, we obtain the following log likelihood:

$$\ln L = -\frac{N}{2}\ln(2\pi|\Sigma_{22}|) - \frac{1}{2}\sum_{i=1}^{N}\begin{pmatrix} y_{1i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta} \\ y_{2i} - \widetilde{\mathbf{X}}_i\boldsymbol{\gamma} \end{pmatrix}^T \Sigma_{22}^{-1}\begin{pmatrix} y_{1i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta} \\ y_{2i} - \widetilde{\mathbf{X}}_i\boldsymbol{\gamma} \end{pmatrix} +$$

$$\sum_{i=1}^{N}\ln\left(\Phi\left(\frac{-d_1 - X_i\delta_2^n - \Sigma_{12}\Sigma_{22}^{-1}\begin{pmatrix} y_{1i} - \widetilde{\mathbf{X}}_i\boldsymbol{\beta} \\ y_{2i} - \widetilde{\mathbf{X}}_i\boldsymbol{\gamma} \end{pmatrix}}{\sqrt{\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}}\right)\Big/\Phi\left(\frac{-d_1 - X_i\delta_2^n}{\sqrt{\Sigma_{11}}}\right)\right)$$

The first part of the log likelihood in equation excluding the last term is identical to the log likelihood of the standard MIMIC model without truncation. The last term corrects for truncation and causes the ML estimates to differ from their least-squares counterparts and ensures that the ML estimates are consistent.

Appendix C: Derivation of the Conditional Predictionof True Sales

Taking the exponential transformation of the structural equation $\ln(S^*) = a + X_1 b + c \ln K_v^* + u$, we obtain $S^* = \exp(a + X_1 b + c \ln K_v^* + u)$, therefore the expectation of true sales conditional on both indicators and causal variables can be rewritten as:

$$\boldsymbol{\mu_{CM}} = \mathbf{E}\left(\mathbf{S}^* | \widetilde{\mathbf{X}}, \mathbf{ln(S^t)}, \mathbf{ln(S^s)}, \mathbf{ln(S^t)} \leq \mathbf{ln(S^*)}\right) =$$

$$\exp\left(a + X_1 b + c\ln(K_v^*)\right) \times E(\exp(u) | \varepsilon_t \leq -d_1 - X\delta_2^n, \mathbf{v}) \qquad (\#1)$$

where $E(\exp(u)|\varepsilon_t \leq -d_1 - X\delta_2^n, \mathbf{v}) = \frac{\int_{-\infty}^{-d_1 - X\delta_2^n} f(\varepsilon_t|\mathbf{v}) \int \exp(u) f(u|\varepsilon_t, \mathbf{v}) du d\varepsilon_t}{\Pr(\varepsilon_t \leq -d_1 - X\delta_2^n|\mathbf{v})}$. Let Φ_1 denote $\Pr(\varepsilon_t \leq -d_1 - X\delta_2^n|\mathbf{v})$. The multivariate normality of $(\varepsilon_t, \mathbf{v})$ implies that $(u|\varepsilon_t, \mathbf{v})$ is normally distributed with the conditional mean and variance as follows:

$$\mu_{u|\varepsilon_t,\mathbf{v}} = E(u|\varepsilon_t, \mathbf{v}) = E(u) + \left(\text{cov}\left(u, (\varepsilon_t, \mathbf{v})\right)\left(\text{var}(\varepsilon_t, \mathbf{v})\right)^{-1} \begin{pmatrix} \varepsilon_t^T \\ \mathbf{v^T} \end{pmatrix}\right)^T \cong$$

$$\begin{pmatrix} \varepsilon_t \\ \ln(S^t) - \widetilde{\mathbf{X}}\boldsymbol{\beta} \\ \ln(S^s) - \widetilde{\mathbf{X}}\boldsymbol{\gamma} \end{pmatrix}^T \Sigma^{-1} \left(\text{cov}\left(u, (\varepsilon_t, \mathbf{v})\right)\right)^T = B_1\varepsilon_t + B_2\left(\ln(S^t) - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) + B_3\left(\ln(S^s) - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\right)$$

and $\sigma^2_{u|\varepsilon_t,\mathbf{v}} = \text{var}(u|\varepsilon_t, \mathbf{v}) = \text{var}(u) + \text{cov}\left(u, (\varepsilon_t, \mathbf{v})\right) \Sigma^{-1} \left(\text{cov}\left(u, (\varepsilon_t, \mathbf{v})\right)\right)^T$,

where $B_i, i = 1,2,3$ denote the three elements of the 3×1 vector $\Sigma^{-1} \left(\text{cov}\left(u, (\varepsilon_t, \mathbf{v})\right)\right)^T$, and let $\mathbf{m}$ denote $\left(B_2\left(\ln(S^t) - \widetilde{\mathbf{X}}\boldsymbol{\beta}\right) + B_3\left(\ln(S^s) - \widetilde{\mathbf{X}}\boldsymbol{\gamma}\right)\right)$. Standardizing $(u|\varepsilon_t, \mathbf{v})$, we have:

$$\int \exp(u) f(u|\varepsilon_t, \mathbf{v})\, du = \int \exp\left(-\frac{\left(u - \mu_{u|\varepsilon_t,\mathbf{v}}\right)^2 - 2\sigma^2_{u|\varepsilon_t,\mathbf{v}}}{2\sigma^2_{u|\varepsilon_t,\mathbf{v}}}\right) \Big/ \sqrt{2\pi\sigma^2_{u|\varepsilon_t,\mathbf{v}}}\, du =$$

$$\int \exp\left(-\frac{\left(u - (\mu_{u|\varepsilon_t,\mathbf{v}} + \sigma^2_{u|\varepsilon_t,\mathbf{v}})\right)^2 - 2\sigma^2_{u|\varepsilon_t,\mathbf{v}}\mu_{u|\varepsilon_t,\mathbf{v}} - \sigma^4_{u|\varepsilon_t,\mathbf{v}}}{2\sigma^2_{u|\varepsilon_t,\mathbf{v}}}\right) \Big/ \sqrt{2\pi\sigma^2_{u|\varepsilon_t,\mathbf{v}}}\, du = \exp\left(\tfrac{1}{2}\sigma^2_{u|\varepsilon_t,\mathbf{v}} + \mu_{u|\varepsilon_t,\mathbf{v}}\right)$$

Plugging the above equation back into the conditional expectation equation (#1) $\overset{\textbf{yield}}{\Longrightarrow}$:

$$\boldsymbol{\mu_{CM}} = \frac{1}{\Phi_1}\exp\left(a + X_1 b + c\ln(K_v^*)\right) \int_{-\infty}^{-d_1 - X\delta_2^n} f(\varepsilon_t|\mathbf{v}) \exp\left(\frac{1}{2}\sigma^2_{u|\varepsilon_t,\mathbf{v}} + \mu_{u|\varepsilon_t,\mathbf{v}}\right) d\varepsilon_t =$$

$$\frac{1}{\Phi_1}\exp\left(a + X_1 b + c\ln(K_v^*) + \tfrac{1}{2}\sigma^2_{u|\varepsilon_t,\mathbf{v}} + m\right) \int_{-\infty}^{-d_1 - X\delta_2^n} f(\varepsilon_t|\mathbf{v}) \exp(B_1\varepsilon_t)\, d\varepsilon_t \qquad (\#2)$$

Next we proceed to derive $\int_{-\infty}^{-d_1 - X\delta_2^n} f(\varepsilon_t|\mathbf{v}) \exp(B_1\varepsilon_t)\, d\varepsilon_t$:

$$\int_{-\infty}^{-d_1-X\delta_2^n} f(\varepsilon_t|\mathbf{v})\exp(B_1\varepsilon_t)\,d\varepsilon_t =$$
$$\int_{-\infty}^{-d_1-X\delta_2^n} \exp\left(-\left(\varepsilon_t-\mu_{\varepsilon_t|\mathbf{v}}\right)^2-\sigma^2_{\varepsilon_t|\mathbf{v}}B_1\varepsilon_t/2\sigma^2_{\varepsilon_t|\mathbf{v}}\right)\Big/\sqrt{2\pi\sigma^2_{\varepsilon_t|\mathbf{v}}}\,d\varepsilon_t =$$
$$\exp\left(\frac{B_1^2\sigma^2_{\varepsilon_t|\mathbf{v}}}{2}+B_1\mu_{\varepsilon_t|\mathbf{v}}\right)\int_{-\infty}^{-d_1-X\delta_2^n}\exp\left(-\left(\varepsilon_t-\left(\mu_{\varepsilon_t|\mathbf{v}}+\sigma^2_{\varepsilon_t|\mathbf{v}}B_1\right)\right)^2/2\sigma^2_{\varepsilon_t|\mathbf{v}}\right)\Big/\sqrt{2\pi\sigma^2_{\varepsilon_t|\mathbf{v}}}\,d\varepsilon_t =$$
$$\exp\left(\frac{B_1^2\sigma^2_{\varepsilon_t|\mathbf{v}}}{2}+B_1\mu_{\varepsilon_t|\mathbf{v}}\right)\Phi\left(\frac{-d_1-X\delta_2^n-\mu_{\varepsilon_t|\mathbf{v}}-\sigma^2_{\varepsilon_t|\mathbf{v}}B_1}{\sqrt{\sigma^2_{\varepsilon_t|\mathbf{v}}}}\right)$$

Plugging the result back into equation (#2) we get the formula for the conditional expectation:

$$\boldsymbol{\mu_{CM}} = \exp\left(a+X_1b+cln(K_v^*)+m+B_1\mu_{\varepsilon_t|\mathbf{v}}+\frac{1}{2}\sigma^2_{u|\varepsilon_t,\mathbf{v}}+\frac{1}{2}B_1^2\sigma^2_{\varepsilon_t|\mathbf{v}}\right)\Phi\left(\frac{-d_1-X\delta_2^n-\mu_{\varepsilon_t|\mathbf{v}}-B_1\sigma^2_{\varepsilon_t|\mathbf{v}}}{\sqrt{\sigma^2_{\varepsilon_t|\mathbf{v}}}}\right)\Big/\Phi_1$$